전자기록물의 메타데이터 추출 및 비교 검증 기술 연구*

Extracting and Validating Metadata in Electronic Records

최 주 호(Joo Ho Choi)**
이 재 영(Jae Young Lee)***

목 차

〈초 록〉

전자기록물의 이관할 때, 전자기록물의 필수 메타데이터의 검증과 실제 문서에 있는 메타데이터를 이용한 검증도 중요하다. 본 연구에서는 전자기록물에 포함된 다양한 형식의 전자파일 중에서 본문파일에서 메타데이터를 추출하고 항목별로 분류한 후 이관되는 메타데이터 항목과 비교 검증을 위한 기술 개발을 연구하였다. 해외에서 개발된 추출 도구와 달리 국내 전자결재 형식을 감안하여 첨부된 본문파일에서 메타데이터를 추출하는 기술을 개발하였으며, 기록물 문서 메타항목에 저장된 원 메타데이터와 추출 메타데이터간 비교 검증을 수행하는 도구를 개발하였다.

주제어: 메타데이터 추출, 메타데이터 검증, 형태소 분석, 파일 필터링, 기술정보은행, 추출 도구, 메타데이터 추출 알고리즘

<ABSTRACT>

When migrate electronic records, the validation of the required metadata in electronic records and verified with the metadata in the document are also important. This paper presents a method and implements a tool to extract data from files in various formats and use them to validate metadata associated with the files in electronic records. Compared to other metadata extraction tools, especially developed in foreign countries, the standard form of documents used in Korean government is taken into account and metadata is extracted from the content of files. The tool compares the extracted data to encapsulated metadata for validation.

Keywords: extraction metadata, file filtering, metadata extraction tool, JHOVE, DFR(Digital Format Registry)

* 본 연구 논문은 "2011년 행정안전부 국가기록원의 기록물 보존기술 연구개발 사업"의 일환으로 진행한 "차세대 전자기록관리 인프라 응용기술 연구 개발" 과제 결과를 토대로 작성됨.
** (주)세미콘네트웍스(iq2chun@gmail.com)
*** (주)세미콘네트웍스(jaeyoung.2ee@gmail.com)
■ 접수일: 2012년 3월 7일 ■ 최초심사일: 2012년 3월 29일 ■ 게재확정일: 2012년 4월 25일

1. 서 론

1.1 연구의 필요성과 목적

차세대 전자기록관리 인프라 기반기술 연구(국가기록원 2010b)는 이관 기록물 검증 단계의 문제점으로서 육안 검수를 지적하였다. 육안검수로 인하여 이관 시간이 지연되므로 대량 이관이 예상되는 2015년 이후에는 큰 문제가 될 것으로 예상하고 있다. 또한 육안검수로 기록물 내용을 검수한다 할지라도 업무담당자가 아니라면 실효성이 적다고 할 수 있다.

즉, 육안검수에 의존하는 현행 검증 기능은 부정확한 검증과 인수 시간 지연의 문제를 일으키므로 자동화된 검증 체계가 논의되었는데 자동화 대상으로는 메타데이터 검수, 전자파일 포맷 검수 등이 제시되었다.

국가기록원에 이관되는 전자기록물의 내용을 점검하고 문서보존포맷 및 원문 파일, 첨부파일이 정상적인지 검사하는 과정을 최대한 자동화함으로써 이관 시간의 현저한 단축과 함께 육안검수 시에는 검사자마다 다른 기준을 적용할 수 있지만 자동화를 통해 일관성 있는 검사기준을 적용할 수 있다. 또한 육안검수 시에는 검증 항목을 누락하는 실수를 할 가능성이 있으나 자동화 시스템을 통해 검증 항목 누락 위험을 제거할 수 있다.

이러한 검증 자동화 과정에는 메타데이터에 대한 검증이 포함되어야 하는데 메타데이터에 대한 검증을 통해 전자기록물의 무결성, 신뢰성, 진본성, 이용가능성이 검증된다. 전자기록물 이관 단계에서 수행하여야 할 메타데이터 검증 항목은 자동누락검사, 중복 검사, 자료 건수, 항목 타입/길이, 데이터 정합성 등에 대한 검수를 진행하여야 한다.

따라서 객관적 검증이 가능한 메타데이터 자동 검증 도구를 개발하여 전자기록물에 대한 신뢰성을 확보하고 기록물 이관 작업의 효율성 확보가 필요하다. 본 연구에서는 2015년 대량으로 인수되는 전자기록물에 대한 메타데이터 검증 방안을 수립하고 테스트베드 도구를 개발하여 그 적용 가능성을 모색하고자 하였다.

1.2 연구의 범위와 방법

본 연구에서는 전자기록물에 포함된 다양한 형식의 전자파일에서 텍스트를 추출하여 형태소 분석 및 색인하고 메타데이터를 추출할 수 있는 기술에 대한 선진 사례를 비교 연구하고 개발을 위한 시사점을 도출한 후 이를 토대로 메타데이터 추출 및 비교 검증 프로세스 수립과 함께 도구를 개발하여 테스트를 함으로써 향후 국가기록원의 전자기록물 이관업무에 편리성을 제공하는 기반을 갖추고자 한다.

2. 국내외 기술 개발 현황

2.1 텍스트 추출 관련 국내외 기술 동향

2.1.1 문헌정보학에서의 정보 추출 연구 동향

의미 기반 검색으로 발전해 가고 있는 가운데 다양한 정보 추출 기법에 대한 연구는 계속되고 있으며, 색인의 정확도와 정보 근접성을 높이기 위한 다양한 색인 기법 연구 또한 지속되고 있다. 정보학과 관련된 학술적 커뮤니케이션의 장

으로서 정보관리에서 정보 기술 응용을 활성화하는 역할을 담당해 온 "정보관리학회지"(한국정보관리학회)에 게재된 연구 논문을 창간해인 1984년부터 2009년까지 25년 간 분석한 연구 결과에 의하면, 논문의 소주제에서 가장 많은 연구는 도서관서비스이며, 그 다음 이용자연구, 자동문헌처리, 도서관통합시스템, 시소러스/온톨로지, 디지털도서관, 웹, MARC/메타데이터, 자동색인/초록 순이다. 연구에서 시소러스/온톨로지, 메타데이터 연구의 증가를 볼 때 의미기반 검색이 화두가 되고 있음을 확인할 수 있다. 정보관리학회지는 문헌정보학 종사자들이 주축이 되므로 컴퓨터 응용 기술 분야에서의 연구까지 고려한다면 이 분야 연구는 더욱 많을 것으로 예상된다.

2.1.2 색인 기술 동향

색인 기술은 색인어 단위에 따라 형태소 단위 색인법과 어절 단위 색인법 그리고 n-Gram 색인법이 사용되고 있다. 대부분의 상용 시스템은 형태소 분석기를 이용하고 있다. 1990년대 초기에는 조사만 분리하거나 10만 단어 수준의 작은 사전을 사용하는 방식으로 형태소 분석기를 이용하거나 bigram 방식을 이용하였다.

2000년부터 인간이 직접 분석한 결과를 이용하는 기분석 사전을 도입하기 시작했으며 높은 검색 성능을 유지하기 위하여 검색 엔진 컨텐츠의 종류(사전, 노래, 뉴스, 커뮤니티, 블로그, 웹문서, 사이트 디렉토리 등)에 따라 다른 속성을 반영하여 사전과 문법을 수정하는 색인기 튜닝이 행해졌다.

색인기 튜닝이 계속적으로 시도되어 뉴스와 같이 증가하는 컨텐츠는 기존 색인을 그대로 유지하면서 새 단어를 사전에 추가하고 다양한 내용의 웹문서 컨텐츠는 기분석 사전을 이용하는 방식이 시도되었다. 2006년 이후 명칭 데이터베이스를 도입하거나 검색 쿼리를 문서에서 추가적으로 추출하는 기법이 도입되었다.

현재 상용 검색 엔진 및 솔루션은 형태소 분석기를 이용하고 있으며 한국어 자연어 처리에 대한 연구 및 한국어 텍스트 분석 품질을 높이기 위한 형태소 분석기에 대한 연구는 지속되고 있다.

2.2 한국어 형태소 분석기 연구 동향

2.2.1 KLT(Korean Language Technology)

국민대학교 강승식 교수에 의해 개발된 KLT는 구 HAM(Hangul Analysis Module)의 이름이 변경되었으며 2010년 10월 5일 version 2.2.0이 출시되어 있으며 현재 가장 많이 사용되고 있는 형태소 분석기이다.

2.2.2 꼬꼬마 형태소 분석기

서울대학교 IDS(Intelligent Data Systems) 연구실에서 개발하였으며 띄어쓰기 오류에 덜 민감한 분석기로 소개되고 있다. 분석 속도 향상을 위해 다양한 최적화 방법을 이용하였으며 분석 품질 향상을 위해 확률 모델을 이용하고 있다.

2.2.3 한나눔 한국어 형태소 분석기

한국과학기술원 시맨틱 웹 첨단연구센터에서 개발하여 관리하고 있는데 오픈 소스로 관리되고 있으며 한나눔 자바 버전 0.83 버전을 다운로드 할 수 있다.

기존 한국어 형태소 분석기들은 대부분 특정

시스템에 맞게 최적화되어 있어 실행 효율성과 정확성을 높이는데 중점을 두고 있으나, 접근성과 확장성이 떨어지고 다양한 요구에 유연하게 대처할 수 없는 단점을 가지고 있는데 이를 해소하기 위하여 플러그인 형태의 형태소 분석 컴포넌트를 개발하였다. 따라서 유연하게 워크플로우를 구성하고 다양한 목적에 맞게 활용할 수 있을 것으로 기대된다. 이 분석기 개발에서 2010년 국가기록원 전자기록물 보존 연구개발 사업의 지원이 있었다.

2.2.4 Maran-CJK

모란소프트에서 개발한 상용 소프트웨어로서 한국어, 일본어, 중국어, 영어 형태소 분석을 수행할 수 있는데 유니코드를 기본으로 제작되었으며 cost 방식에 의해 추출하는 특징이 있다.

2.2.5 moHANA

워드워즈에서 개발한 상용 소프트웨어로서 다차원 해석 사전을 기반으로 하여 태그 정보 사전, 어휘 사전, 문법 사전을 추가적으로 이용하는 특징이 있다.

2.3 해외 정보 추출 연구 동향

2.3.1 Text Analysis Conference

1) Message Understanding Conferences (MUC)

미국 Defense Advanced Research Projects Agency(DARPA)의 지원을 받아 Naval Ocean Systems Center(NOSC)가 군사 텍스트 분석 기술의 연구를 촉진하기 위하여 정보 추출 기술 평가와 평가 결과를 공유하는 컨퍼런스를 개최하기 시작하였다. 1987년 MUC-1 개최 후 1997년 MUC-7까지 7회의 컨퍼런스가 개최되었으며 이 과정에서 추출 과제에 Named Entity와 Conference가 추가되고 더욱 정교해졌다.

2) Automatic content Extraction(ACE)

MUC의 다음 단계로서 1999년부터 2008년까지 NIST(National Institute of Standards and Technology)는 ACE 프로그램으로서 평가와 컨퍼런스를 열었으며 여기서 뉴스, 회의 녹취록, 웹로그 등 7개 분야에서 선정된 문서를 대상으로 문서 내에서 엔터티와 관계 추출, 다른 문서와 상호 참조 관계로서 전역적 엔터티와 과제 추출의 4개 과제로 평가하였다.

3) Text Analysis Conference(TAC)

2009년부터 ACE는 TAC 트랙으로 구성되어 진행되어 오고 있다. 이 프로그램에 IBM Watson Research Center, BBN Technologies, Cortex Intelligence 등 미국, 유럽 및 중국의 기관과 대학이 참가하고 있다.

2011년 TAC는 Knowledge Base Population, Textual Entailment Recognition, Summarization 3가지 부문으로 나뉘어 진행될 예정이다.

Knowledge Base는 외부의 지식 소스를 참조하여 엔터티에 대한 정보 추출, 온톨로지에 인명, 조직, 위치 데이터를 생성하고 텍스트에서 추출한 정보를 추가하는 기능 평가를 수행한다.

Textual Entailment Recognition은 한 텍스트로부터 추론할 수 있는 다른 텍스트를 추출하는 기능으로 새로 출현한 단어를 발견하는 기능이다. Knowledge Base Population 과제의 산

출물을 대상으로 관계가 있는지 검증하는 기능을 평가한다.

마지막으로 Summarization은 텍스트로부터 의미 있는 요약을 생성하는 기능을 평가한다.

2.3.2 GATE(General Architecture for Text Engineering)

1) GATE 개요

University of Sheffield의 1995년 프로젝트로부터 시작된 GATE는 오픈 소스로 개발되어 왔으며 GNU LGPL 라이센스 하에서 상용 및 연구 목적으로 사용할 수 있다. 언론, 제약 및 암 연구, 웹 마이닝, 웹 아카이빙 등 여러 분야의 다양한 시스템에서 텍스트 분석, 정보 추출, 시맨틱 어노테이션을 위해 활용되고 있다. UMC, ACE 등의 평가 프로그램에 참가하여 검증을 받았다.

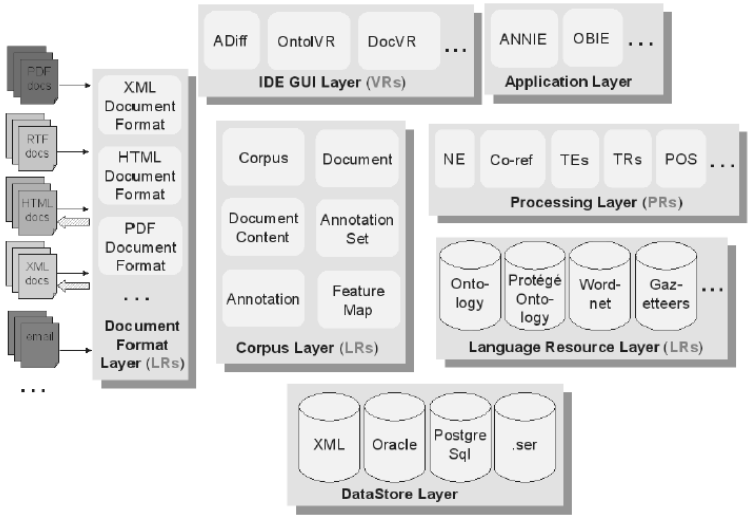
2) GATE 시스템 구성

〈그림 1〉에서와 같이 GATE 시스템은 Java로 개발되었으며 Linux, Windows, MacOS에서 테스트되었으며 2011년 5월 현재 Release 6.1이 제공되고 있다.

GATE는 GATE Developer, GATE Embedded, GATE Teamwear 제품군으로 구성된다. GATE Developer는 정보 추출 시스템 및 약 50여개 플러그인과 통합되는 데스크탑 개발 환경으로 ANNIE와 시맨틱 어노테이션 도구가 플러그인으로 포함되어 있다. GATE Embedded는 다른 애플리케이션 내에서 임베드되어 GATE Developer가 사용하는 서비스에 접근할 수 있도록 하는 객체 라이브러리이다. GATE Teamware는 협업적 어노테이션 환경을 제공한다.

3) ANNIE(A Nearly-New Information Extraction System)

ANNIE는 GATE의 정보 추출 시스템으로 유한 상태 알고리즘과 JAPE(Java Annotation



〈그림 1〉 GATE 시스템 구성도(GATE system structure)

Patterns Engine)을 기반으로 개발되었다.

ANNIE를 구성하는 컴포넌트는 Document Reset, Tokeniser, Gazetteer, Sentence Splitter, RegEx Sentence Splitter, Part-of-Speech Tagger, Semantic Tagger, Orthographic Coreference(OrthMatcher), Pronominal Coreference로 구성되어 있다.

컴포넌트에 대하여 간략히 설명하면 Document Reset은 입력 문서에서 본래 태그 외의 태그를 제외하고 컨텐츠만 남게 정리한다. Tokeniser는 텍스트를 숫자, 구두점 등으로 분리한다. Gazetteer은 엔터티 목록에 있는 엔터티 이름을 추출한다. Setntence splitter는 텍스트를 문장으로 분리하는 역할을 한다. RegEx Sentence Splitter는 Sentence Splitter를 보완하여 정규식에 기반하여 텍스트를 문장으로 분리한다. Part-of-Speech Tagger는 Brill tagger를 변형한 태커로서 각 단어나 특수문자에 태그를 붙이며 Semantic Tagger는 주석이 붙은 엔터티를 생성하며 Orthographic Coreference(OrthoMatcher)는 발견된 엔터티 사이의 관계 분석을 수행하며 마지막으로 Pronominal Coreference는 대용어 분석 기능을 수행한다.

3. 메타데이터 검증 기술 연구

본 연구에서는 장기보존포맷 규격[1] 상의 메타데이터 요건을 이해하고 원문 파일을 이용하여 검증 가능한 메타데이터를 분석하였으며, 원문 전자 파일로부터 메타데이터를 추출하는 방법을 연구하여 제시하고자 한다.

이를 위해 국내외 메타데이터 규격을 검토하였고 장기보존포맷을 비롯하여 전자기록물 영구보존을 위한 국가기록원 규격[2] 및 장기보존을 위한 메타데이터 규격과 전자기록물 검증 관련 해외 규격을 분석 검토하였다. 또한 인수 대상 전자기록물에 대한 현황 분석과 함께 전자파일로부터 메타데이터를 추출 하는 기술에 대하여 검토하였다.

마지막으로 원문 텍스트 활용 체계 수립을 위하여 형태소 분석을 기반으로 주제어를 추출하여 적용하는 각종 기술의 선행 연구 결과를 검토하고 적용 방안을 연구하였다.

3.1 메타데이터 추출 도구 비교 연구

파일로부터 메타데이터를 추출하는 대표적인 도구로서 뉴질랜드 국립도서관 Metadata Extraction Tool, JHOVE, 국가기록원 DFR을 비교 검토하였는데, 비교 대상인 3개 도구를 개략적으로 비교한 내용은 〈표 1〉 추출 도구 비교에 나타내었다. 뉴질랜드 국립도서관의 Metadata Extraction Tool은 단독 추출 도구이며, JHOVE는 식별 도구와 함께 제공되는데 비해 국가기록원 DFR은 레지스트리의 기능만을 제공하도록 구성되어 있다.

도구별로 지원 포맷에 차이가 있으나 문서 포맷 중 PDF는 모두 지원하고 있다. PDF 버전에 대해서 JHOVE는 버전을 명시하고 있으나 다

1) NAK-TS_1-2_2008_기록관리시스템과_영구기록관리시스템간_데이터_연계규격.pdf
2) 기록관리시스템 데이터연계 기술규격 제1부 업무관리시스템과의 연계.pdf
기록관리시스템 데이터연계 기술규격 제3부 기능분류시스템과의 연계.pdf

〈표 1〉 추출 도구 비교(Comparison of extraction tools)

	Metadata Extraction Tool	JHOVE	DFR
개발 주체	뉴질랜드 국립도서관	California Digital Library (CDL), Portico, Stanford University * JHOVE1.0은 JSTOR와 하버드 대학교 도서관	대한민국 국가기록원
라이센스	공개 소스	공개 소스	비공개
지원 포맷	약 20종	JHOVE2.0 약 10종	약 18종
포맷 지원 방법	포맷별 모듈화	포맷별 모듈화	포맷별 모듈화
실행 방법	단독 도구	식별, 검증, 평가 도구와 함께 제공	레지스트리, 식별, 검증 도구와 함께 제공
실행 방법	그래픽인터페이스 명령어 라인	명령어 라인	웹서비스

〈표 2〉 추출 도구별 지원 포맷(Comparison of the formats supported by extraction tools)

	Metadata Extraction Tool	JHOVE2[3)]	DFR
문서	MS Word(version 2, 6) Word Perfect Open Office(version 1) MS Works MS Excel MS PowerPoint PDF	PDF * PDF 1 - 1.7, ISO 32000-1, PDF/X-1(ISO 15930-1), PDF/X-1(ISO 15920-1), \-1a (ISO15930-4), \-2(ISO 15930-5), \-3(ISO 15930-6), PDF/A-1(ISO 19005-1)	MS Word MS Excel MS PowerPoint 한컴오피스 훈민정음 아리랑 하나워드
이미지	BMP GIF JPEG TIFF	ICC color profile JPEG 2000 : JP2(ISO/IEC 15444-1), JPX(ISO/IEC 15444-2) TIFF : 4 - 6, Class B, G, R, P, Y, TIFF/IT, TIFF/EP, Exif, GeoTIFF, DNG	GIF JPEG TIFF
음성	WAVE MP3 BFW FLAC	WAVE : Broadcast Wave Format(EBU N22-1997)	AIFF WAVE
텍스트	HTML XML	SGML XML UTF-8 : ASCII(ANSI X3.4)	ASCII HTML XML UTF-8
기타	ARC	ZIP Shapefile GZIP(예정) ARC(예정)	

른 도구들은 PDF로만 표시하고 있다. 마이크로소프트 오피스 문서는 JHOVE에서는 지원되지 않으나 Shapefile, ZIP, ICC 프로파일이 지원된다. 국가기록원 DFR은 국내 한컴오피스, 훈민

3) JHOVE2는 JHOVE1의 후속 버전이나 지원 포맷이 다름. 아래 지원 포맷 중 2011년 4월 공개된 v2.0.0에서 지원되지 않고 있는 포맷도 있으며 향후 버전에서 지원될 예정임.

정음 등의 포맷이 지원되는 특징을 갖고 있다.

3.2 도구별 추출 데이터 비교 연구

3.2.1 뉴질랜드 국립도서관 Metadata Extraction Tool

뉴질랜드 국립도서관이 2002년부터 개발하여 활용하고 있는 보존 메타데이터 스키마에서 정한 파일 메타데이터 항목을 추출하는데 〈표 3〉에 뉴질랜드 국립도서관 지정 파일 보존 메타데이터 항목을 나타내었다.

파일 크기, 생성 일시, 포맷 등 파일 헤더에서 얻을 수 있는 데이터와 문서 파일, 이미지, 음성, 동영상에 따라 각각 필요한 특성 정보를

〈표 3〉 뉴질랜드 국립도서관 지정 파일 보존 메타데이터
(Preservation metadata designated by National Library of New Zeland)

종류	항목명	설명	예제	필수
공통	객체 ID	뉴질랜드 국립도서관 내에서 디지털 객체에 부여하여 관리하는 식별 번호	875	Y
	파일 ID	뉴질랜드 국립도서관 내에서 복합 객체 내 파일에 부여하여 관리하는 식별 번호	34-1, 34-2	Y
	파일 경로	복합 객체 내 파일의 위치 정보	…/birds-of-new-zealand/bird-songs/kiwi.wav	N
	파일명과 확장자	파일 이름과 확장자	98_pm_01.doc	Y
	입수 파일명	최초 입수 당시의 파일명	File & Folder Utilities version 1.doc	N
	파일 크기	파일 크기	200.6GB	Y
	생성 일시	파일 생성 일자와 시각	2002-04-18 14:32:51	N
	MIME Type	파일의 MIME 타입	image/gif application/msword image/x-cdc	Y
	포맷명	포맷 문서에 명시된 공식 이름	MS Word 2000, MPEG	Y
	포맷 버전	포맷의 버전	XP, V2.0	N
	타겟 표시자	복합 객체에서 객체 전체에 접근하게 해주는 객체 내 파일이 있는지 여부	YES 또는 No	N
텍스트	문자 세트	적용된 문자 세트	ASCII, Unicode, EBCDIC, UTF-8	Y
	마크업 언어	텍스트를 마크업하는데 사용된 언어	SGML, XML, HTML	Y
이미지	해상도	이미지 해상도	600 dpi; 320 dpi, 1500 d/cm	Y
	크기	픽셀 수로 표시하는 가로 세로 크기	4096 x 6144 pixels	Y
	샘플 당 비트	각 픽셀에 대하여 컴포넌트 당 비트 수	1 = 1 bit(bitonal) 4 = 4 bit grayscale 8 = 8 bit grayscale or palletised colour 8,8,8 = RGB 16,16,16 = TIFF, HDR (high dynamic range) 8,8,8,8 = CMYK	Y
	컬러 스페이스	비압축 이미지 데이터에 대한 컬러 스페이스 지정	0, 1, 2, 3, 4, 5, 6, 7, 8	Y

종류	항목명	설명	예제	필수
이미지	ICC 프로파일명	ICC(International Color Consortium) 프로파일명	PhotoCD: OptiCal: Profile/80: Softproof (Photoshop plug-in)	Y
	컬러맵 위치	컬러맵 파일의 위치	URL	Y
	방향	디스크에 저장된 이미지의 방향	1 = normal*, 3 = normal rotated 180°, 6 = normal rotated cw 90°, 8 = normal rotated ccw 90°, 9 = unknown	Y
	압축 방법	압축 종류와 압축률	4 = ITU Group 4	Y
음성	Resolution	음성 파일 생성 시 적용된 샘플 레이트	32100, 44100, 192000	Y
	Duration	음성 파일의 실행 시간	01:27:38:247	Y
	Bit Rate	음성 엔코딩에 사용된 워드 길이	16, 20, 24	Y
	Compression	압축 방법	MPEG 3, Dolby A	Y
	Encapsulation	파일의 배포 포맷명과 버전	Real Audio II	Y
	Channels	채널 수와 관계를 식별하는 소리 포맷	Mono 2 channel stereo 5 channel surround	Y
비디오	Frame Dimension	단일 정지 화상 프레임의 해상도	640 pixels x 480 pixels	Y
	Duration	동영상 파일의 실행 시간	01:27:38:247	Y
	프레임 수	동영상의 프레임 수	10000	Y
	Frame Rate	초당 프레임 수(fps)	25	Y
	코덱	적용된 코덱명, 버전	DivX 5.0.5	Y
	화면 비율	요구되는 화면 비율	4:3	Y
	스캔 모드	progressive/interlaced	progressive, interlaced	Y
	음성	동영상 내 음성 포함 여부	Yes, No	Y

추출한다. 특히 이미지, 음성, 동영상 포맷에 대하여 재생을 위한 상세 정보를 추출한다. 규격 스키마에 맞춘 XML 파일을 생성하여 추출되지 않는 데이터는 공백 엘리먼트로 출력된다.

뉴질랜드 국립도서관의 Metadata Extraction Tool을 이용하여 지원되지 않는 한글 HWP 포맷 파일에서 메타데이터를 추출하면 〈그림 2〉와 같이 파일 형식, 생성 일시만 추출된다.

```
<File xmlns:nz_govt_natlib_xsl_XSLTFunctions="nz.govt.natlib.xsl.XSLTFunctions">
<FileIdentifier>0-2</FileIdentifier>
<Filename>
<Name>hwpv30.hwp</Name>
<Extension>hwp</Extension>
</Filename>
<FormerFilename>
<Name/>
<Extension>hwp</Extension>
</FormerFilename>
<Size>9000</Size>
<FileDateTime>
<Date format="yyyyMMdd">20111101</Date>
<Time format="HHmmssSSS">151949078</Time>
</FileDateTime>
<Mimetype>file/unknown</Mimetype>
</File>
```

〈그림 2〉 HWP에서 추출 결과(Extraction results against HWP format)

3.2.2 JHOVE

JHOVE[4]에서 지원되지 않는 포맷은 BaseFormatModule에 의해 기본 사항만 추출되며 여기에는 포맷, 포맷 버전, 파일 크기, 최종 수정 일시가 포함된다. JHOVE에서 지원하는 포맷 파일은 포맷의 고유 속성이 추출된다. JHOVE2 v2.0.0에서 지원되는 TIFF 포맷 파일로부터 메타데이터를 추출하면 〈표 4〉와 같은 포맷 속성 값이 추출된다.

PDF 포맷의 경우 JHOVE2 v2.0.0에서는 모듈이 지원되지 않으며 JHOVE1.0을 이용하여 추출하면 생성 소프트웨어, 생성일, 최종 수정일, 페이지 레이아웃, 폰트 등이 추출되는데 〈그림 3〉은 JHOVE1에서 PDF 파일의 메타데

〈표 4〉 TIFF 포맷 속성(TIFF format attributes)

속성 클래스	설명
IFH	IFH(Image File Header)의 속성 추출 ByteOffset, ByteOrdering, FirstIFD, MagicNumber
IFD	IFD(Image File Directgory Entry)에 대한 속성 추출 IFD 엔트리 수, IDF의 바이트 오프세트
IFDEntry	각 IFD(Image File Directgory Entry) 내 속성 추출 압축 방법, 엔트리 수 등
TiffIFD	TIFF Version 6.0 IFD 속성 추출

〈그림 3〉 JHOVE1의 PDF 포맷 메타데이터 추출 결과
(PDF format metadata extractd by JHOVE1)

4) 〈https://bytebucket.org/jhove2/main/wiki/documents/JHOVE2-functional-requirements〉.

이터를 추출한 결과 화면이다.

3.2.3 국가기록원 DFR

국가기록원 DFR은 포맷별로 각기 다른 모듈이 포맷의 특성에 기반하여 데이터를 추출하기 때문에 포맷에 따라 데이터 항목이 매우 다르다. 국가기록원의 DFR은 문서 파일 포맷 특히 다른 도구들과 달리 국내의 지배적인 문서 포맷에서 데이터를 추출할 수 있기 때문에 국내 포맷으로부터 추출 가능한 데이터에 대한 비교를 〈표 5〉에 나타내었다.

국가기록원 DFR에서 지원하는 국내 파일 포맷에서 공통으로 추출 가능한 데이터는 포맷, 포맷 버전, 파일 크기, 생성 일시, 최종 수정 일시이며 하나워드와 훈민정음 포맷을 제외하면 저자와 주제 또는 주제어 추출이 가능하다.

국내 포맷 이외의 텍스트, 이미지, 음성 파일 포맷인 경우에는 포맷, 포맷 버전, 파일 크기, 생성 일시를 모두 추출하며 그 외 문서 위치를 나타내는 URI와 포맷별 고유 속성을 추출하는데 고유 속성은 JHOVE 1.0에서 사용되는 모듈을 이용하므로 JHOVE 1.0과 동일한 데이터를 추출한다.

3.3 추출 도구 구조 및 추출 방법 비교 연구

3.3.1 추출 도구 구조 연구

뉴질랜드 Metadata Extraction Tool과 JHOVE2 (및 JHOVE1) 및 국가기록원 DFR 모두 포맷에 따라 별도로 실행되는 모듈을 이용

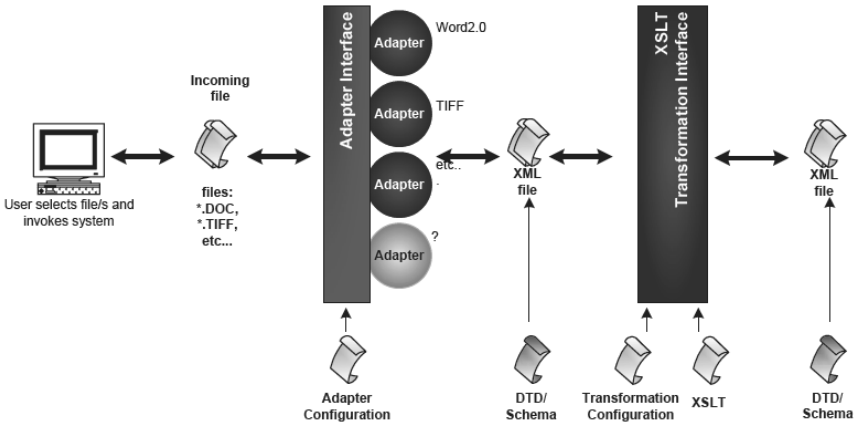
〈표 5〉 국가기록원 DFR의 추출 메타데이터(Metadata extracted by National Archives DFR)

	한컴오피스 HWP	한컴오피스 슬라이드	한컴오피스 넥셀	MS 오피스 (doc, ppt, xls)	아리랑	하나워드	훈민정음
포맷	○	○	○	○	○	○	○
포맷 버전	○	○	○	○	○	○	○
파일 크기	○	○	○	○	○	○	○
생성 일시	○	○	○	○	○	○	○
최종 수정 일시	○	○	○	○	○	○	○
제목	○	○	○	○	○	○	
주제	○	○	○	○			
주제어				○	○		
저자	○	○	○	○	○		
문서 요약	○						
미리보기 텍스트	○						
관리자		○					
회사		○					
페이지 수		○	○	○	○	○	
단어 수			○	○			
문자 수			○	○			
메모		○	○	○			
템플릿			○	○			
로고					○		

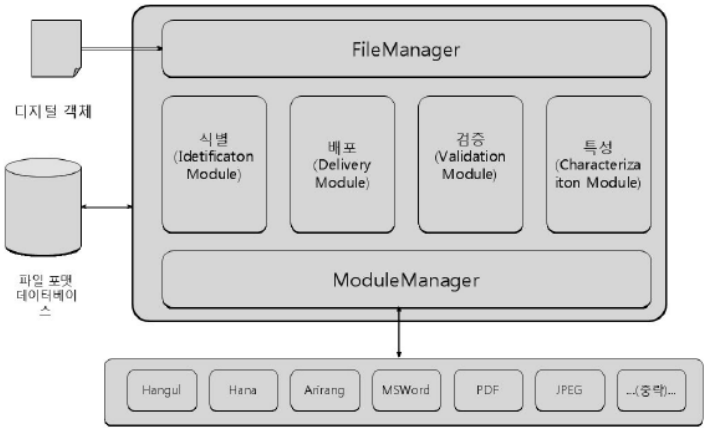
하는 구조를 취하고 있다. 〈그림 4〉 및 〈그림 5〉와 같이 어댑터, 모듈 등의 용어로 각각의 포맷을 처리하는 프로그램을 추가, 수정이 용이하도록 구성하고 있다. 이처럼 포맷별로 모듈화해야 하는 이유는 포맷별 추출 가능한 속성과 방법이 매우 상이하다는 점과 포맷에 따라 서로 다른 개발 주체의 소스 유지관리가 용이하며 지원되는 포맷의 추가 및 수정이 편리하기 때문이다.

3.3.2 내장 메타데이터 이용

포맷에서 특성 정보를 추출하려면 우선은 포맷이 식별되어야 하기 때문에 포맷 종류와 포맷 버전은 모든 도구에서 공통적으로 추출한다.



〈그림 4〉 뉴질랜드 도서관 Metadata Extraction Tool 아키텍처
(National Library of New Zealand Metadata Extraction Tool Architecture)



〈그림 5〉 국가기록원 DFR 포맷 특성 추출 아키텍처
(National Archives DFR format attributes extraction architecture)

또한 파일의 최종 저장일과 크기는 추출되는 속성 중 모든 도구 및 포맷에서 공통적으로 추출된다.

그러나 도구나 포맷에 따라 추가적으로 속성을 추출하기 위해서 포맷에 따라 내장된 메타데이터를 찾아 이용할 수 있다. 〈그림 6〉에서와 같이 PDF 포맷은 눈으로 식별할 수 없는 비트 스트림 내에 일반 텍스트처럼 포함되어 있는 메타데이터가 존재한다. 〈그림 6〉은 PDF 파일을 텍스트 편집 화면으로 오픈한 화면인데 생산한 소프트웨어가 ezPDFBuilder이며 생성일, 수정일 등을 식별할 수 있으며 이 정보를 PDF의 특성 정보로 추출할 수 있다.

한글 HWP 포맷의 경우 〈그림 7〉과 같이 특

〈그림 6〉 PDF 포맷 내장 메타데이터(PDF format built-in metadata)

설명	길이 (바이트)	압축
파일 인식 정보	30	
문서 정보	128	
문서 요약	1008	
정보 블록 (#0)	가변	
글꼴 이름	가변	√
스타일	가변	√
문단 리스트	가변	√
추가 정보 블록 (#1)	가변	√
추가 정보 블록 (#2)	가변	

오프셋	자료형	길이 (바이트)	의미
0	hchar array[56]	112	제목
112	hchar array[56]	112	주제
224	hchar array[56]	112	지은이
336	hchar array[56]	112	날짜
448	hchar array[2][56]	112×2	키워드
672	hchar array[3][56]	112×3	기타
전체 길이		1008	

〈그림 7〉 HWP V3.0의 문서 구조(Document structure for HWP V3.0)

정한 문서 구조를 가지며 그 구조 속에서 본문과 다른 위치에 문서 요약 데이터를 가지고 있다. 이 데이터는 한글 편집기 내 파일 메뉴 하위의 [문서 정보] 메뉴를 통해 입력된 값으로서 국가기록원 DFR이 HWP 포맷에서 추출하고 있는 속성이다.

4. 연구 및 개발 결과

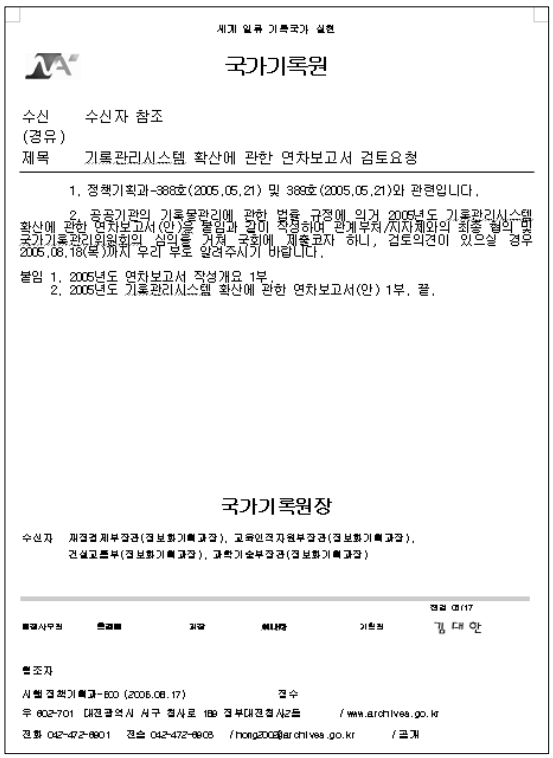
4.1 본문으로부터 메타데이터 추출의 필요성

본 연구에서는 국외 연구 사례처럼 비트스트림에서 메타데이터를 추출하는 방식이 아니라, 본문에서 메타데이터를 추출하는 방안에 대한 연구를 수행하였다.

이는 국내 공공기관에서 생산된 전자기록물의 대다수가 전자문서시스템이나 업무관리시스템을 통하여 생산된 문서이며 특히 결재 문서가 대다수를 차지하고 있으므로 결재 문서의 특성을 활용하는 방안이 필요하다는 판단이었으며, 이를 위해 본문에서 메타데이터를 추출하는 방안에 대한 연구를 수행하였다.

4.2 원문 구성에 대한 검토

〈그림 8〉과 같이 전자문서시스템이나 업무관리시스템에서 생성된 결재 문서는 일정한 양



세계 일류 기록국가 실현

국가기록원

수신 수신자 참조
(경유)
제목 기록관리시스템 확산에 관한 연차보고서 검토요청

1. 정책기획과-388호(2005.05.21) 및 389호(2005.05.21)와 관련입니다.

2. 공공기관의 기록물관리에 관한 법률 규정에 의거 2005년도 기록관리시스템 확산에 관한 연차보고서(안)을 붙임과 같이 작성하여 관계부처/지자체와의 최종 협의 및 국가기록관리위원회의 심의를 거쳐 국회에 제출코자 하니, 검토의견이 있으실 경우 2005.08.18(목)까지 우리 부로 알려주시기 바랍니다.

붙임 1. 2005년도 연차보고서 작성개요 1부.
2. 2005년도 기록관리시스템 확산에 관한 연차보고서(안) 1부. 끝.

국가기록원장

수신자 재정경제부장관(정보화기획과장), 교육인적자원부장관(정보화기획과장),
건설교통부(정보화기획과장), 과학기술부장관(정보화기획과장)

협조자

시행 정책기획과-800 (2005.08.17) 경수

우 602-701 대전광역시 서구 청사로 189 정부대전청사2동 / www.archives.go.kr

전화 042-472-8801 전송 042-472-8905 / hong2008@archives.go.kr / 홍개

〈그림 8〉 전자문서시스템에서 생산된 결재 문서

〈그림 9〉 결재 문서 파일의 문서 정보

식에 제목, 수신자, 기안자 및 결재자 서명, 문서 번호, 시행일/접수일, 연락처 등이 기입되고 기안의 목적, 시행 내용 등이 최대한 요약된 본문으로 구성된다. 〈그림 8〉의 예시 문서에서 제목, 발신인, 결재선, 결재일자, 시행일자, 연락처, 공개구분, 수신자 정보를 찾을 수 있다.

그러나 HWP 포맷의 파일인 예시 문서에 대한 문서 정보(비트스트림)를 조회해 보면 〈그림 9〉와 같이 제목, 지은이에는 부정확한 데이터가 저장되어 있고 주제 등에는 공란으로 데이터가 저장되어 있지 않음을 알 수 있다.

이는 전자문서시스템에서 기안문이 생성될 때 문서 요약 정보가 자동 입력되지 않기 때문에 전자문서시스템이 제공하는 결재문서 편집기를 통해 입력된 데이터를 결재 양식 템플릿 내 정해진 위치에 삽입하는 방식으로 구현되기 때문이다. 그 결과 템플릿의 문서 정보가 복사되거나 존재하지 않게 된다.

따라서 문서 정보에 저장된 데이터는 추출하여도 장기 보존을 위한 메타데이터로 활용하기는 어렵다고 판단된다. 즉 한글 HWP를 비롯하여 국내에서만 존재하는 문서 포맷 파일은 포맷 고유의 구조에 저장된 데이터를 기반으로 메타데이터를 추출하여도 그 활용 효과는 극히 낮다고 볼 수 있다.

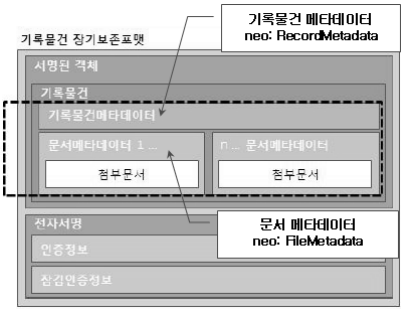
4.3 검증 대상 메타데이터에 대한 연구

결재 문서 원문으로부터 추출한 메타데이터를 비교 검증할 메타데이터는 기록물건 또는 파일에 대한 메타데이터이며, 기록관리시스템으로부터 이관되는 연계 데이터 파일 중 〈표 6〉과 같다. 즉 텍스트 형식 파일 4개와 장기보존 포맷 파일 (.neo) 내에 캡슐화되어 있는 2개 이상의 파일 내에 작성되어 있는 메타데이터를 대상으로 검증하여야 한다.

검증 대상 파일의 메타데이터 중 결재 문서 원문의 데이터 필드와 일치하는 항목을 추출하면 〈표 7〉에서와 같이 문서 제목, 시행/접수 일자, 행위자(기안자, 검토자, 결재자, 협조자) 정보, 수발신자, 공개구분 정보 등이 있다.

기록물건 메타데이터로 이관되는 메타데이터 파일별로 검증 가능한 데이터 항목을 정리하면 〈표 8〉과 같이 2~3개 항목에 대한 검증

〈표 6〉 기록물건 메타데이터 검증 대상(Metadata to be validated for electronic records)

이관 대상	파일명	검증 대상
기록물철 이관 메타데이터	기관코드_TRANS_MST_{$TimeStamp}.txt	N
기록물철 이관인수인계 메타데이터	기관코드_TRANS_TAKE_{$TimeStamp}.txt	N
기록물건 이관메타데이터	**기관코드_TRANS_DTL_{$TimeStamp}.txt**	**Y**
기록물철 행위자 메타데이터	기관코드_TRANS_MPROD_{$TimeStamp}.txt	N
기록물건 행위자 메타데이터	**기관코드_TRANS_DPROD_{$TimeStamp}.txt**	**Y**
기록물철 관계 메타데이터	기관코드_TRANS_MREF_{$TimeStamp}.txt	N
기록물건 관계 메타데이터	**기관코드_TRANS_DREF_{$TimeStamp}.txt**	**Y**
문서 메타데이터	**기관코드_TRANS_DFILE_{$TimeStamp}.txt**	**Y**
장기보존문서	**/neo/DOC+대표기관코드(7)+년도(4)+일련번호(10).neo**  	**Y**
종료파일	기관코드_TRANS_END_{$TimeStamp}.inf	N

〈표 7〉 일반기안문 데이터 항목과 기록물건 메타데이터 비교

데이터 필드		설명	기록물건 이관 메타데이터 종류	데이터 항목
두문	기관명	생산기관명	장기보존〉기록물건 메타데이터	행위자〉기관〉기관명
	수신자	시행문의 수신처 수신처가 2곳 이상이면 "(수신처 참조)" 표시 후 발신인명 아래 기입	기록물건 이관	수발신자
			장기보존〉기록물건 메타데이터	생산정보〉수발신정보〉수신자
	제목	문서 제목	기록물건 이관	공식표제
			장기보존〉기록물건 메타데이터	표제〉공식표제
			장기보존〉문서 메타데이터	* 문서유형= "본문"일 때 문서제목
결문	발신명의	발신자명	기록물건 이관	수발신자
			장기보존〉기록물건 메타데이터	생산정보〉수발신정보〉발신자
	수신자	수신처가 2곳 이상 시 작성	기록물건 이관	수발신자
			장기보존〉기록물건 메타데이터	생산정보〉수발신정보〉수신자
	기안자	기안자 직위	기록물건 행위자	* 행위자유형코드= "기안자"일 때 행위자직위직급명
			장기보존〉기록물건 메타데이터	행위자유형=기안자 행위자〉개인〉직위명

데이터 필드		설명	기록물건 이관 메타데이터 종류	데이터 항목
결문	중간검토자	검토자 직위	기록물건 행위자	* 행위자유형코드=“검토자”일 때 행위자직위직급명
			장기보존>기록물건 메타데이터	* 행위자유형코드=“검토자”일 때 행위자>개인>직위명
	결재자	결재자 직위	기록물건 행위자	* 행위자유형코드=“결재자”일 때 행위자직위직급명
			장기보존>기록물건 메타데이터	* 행위자유형코드=“결재자”일 때 행위자>개인>직위명
	협조자	협조자 직위	기록물건 행위자	* 행위자유형코드=“협조자”일 때 행위자직위직급명
			장기보존>기록물건 메타데이터	* 행위자유형코드=“협조자”일 때 행위자>개인>직위명
	서명[5]	기안자, 검토자, 결재자, 협조자의 전자서명 또는 이미지 서명[주1]	기록물건 행위자	행위자명
			장기보존>기록물건 메타데이터	행위자>개인>개인명
			장기보존>문서 메타데이터	* 문서유형=“본문”일 때 문서행위자
	문서번호	생산부서-일련번호	기록물건 이관	[6]생산(접수)등록번호 13자리 중 뒤 6자리 일련번호[주2]
			기록물건 행위자	
			기록물건 관계	
			문서	
			장기보존>기록물건 메타데이터	고유식별자>기록물건식별자
	시행/접수 일자	(YYYY.MM.DD)	기록물건 이관	생산(접수)일시
			기록물건 행위자	
			기록물건 관계	
			문서	
			장기보존>기록물건 메타데이터	생산정보>수발신정보>시행일자
			장기보존>문서 메타데이터	문서일시
	주소	기안자/처리과/기관 주소	장기보존>기록물건 메타데이터	* 행위자유형코드=“기안자”일 때 행위자>개인>주소
	홈페이지	기안자/처리과/기관 홈페이지	-	-
	전화	기안자/처리과/기관 전화번호	-	-
	팩스	기안자/처리과/기관 팩스	-	-
	전자우편	기안자/처리과/기관전자우편	장기보존>기록물건 메타데이터	* 행위자유형코드=“기안자”일 때 행위자>개인>이메일
	공개구분	공개/부분공개(#)/ 비공개(#) * #: 1개 이상 비공개 사유 번호	기록물건 이관	공개여부 비공개사유
			장기보존>기록물건 메타데이터	권한정보>공개>공개 여부 권한정보>공개>비공개사유

5) 서명이 전자서명일 경우 문자열 데이터로서 추출이 가능하나 이미지 서명은 이름 추출 불가.

6) 생산(접수)등록번호는 결재 문서 내에 표시하는 양식과 이관 연계 파일 내의 메타데이터 값으로 전달되는 형식이 다르다. 메타데이터 값으로는 처리과코드(XXXXXXX)와 일련번호(NNNNNN)를 이어 13자리 문자열(XXXXXXXNNNNNN) 형식을 취하나, 결재 문서에는 인간이 이해 가능하도록 처리과코드 대신 처리과명을 사용하여 “처리과명-일련번호”의 형식으로 표기한다.

〈표 8〉 이관 메타데이터별 검증 가능 항목

이관 대상	파일명	검증 대상
기록물건 이관메타데이터	기관코드_TRANS_DTL_{$TimeStamp}.txt	생산(접수)등록번호, 생산(접수)일시 공식표제, 수발신자, 공개여부, 비공개사유
기록물건 행위자 메타데이터	기관코드_TRANS_DPROD_{$TimeStamp}.txt	생산(접수)등록번호, 생산(접수)일시 * 행위자유형코드 =기안자/검토자/결재자/ 협조자일 때 행위직위직급명, 행위자명[주2]
기록물건 관계 메타데이터	기관코드_TRANS_DREF_{$TimeStamp}.txt	생산(접수)등록번호, 생산(접수)일시
문서 메타데이터	기관코드_TRANS_DFILE_{$TimeStamp}.txt	생산(접수)등록번호, 생산(접수)일시
장기보존문서	/neo/DOC+대표기관코드(7)+년도(4)+일련번호(10).neo 〉 RecordMetadata	고유식별자〉기록물건식별자 권한정보〉공개〉공개 여부 권한정보〉공개〉비공개사유 생산정보〉수발신정보〉발신자 생산정보〉수발신정보〉수신자 생산정보〉수발신정보〉시행일자 표제〉공식표제 * 행위자유형=기안자 일 때 행위자〉기관〉기관명 행위자〉개인〉개인명[주2] 행위자〉개인〉이메일 행위자〉개인〉주소 * 행위자유형= “검토자/결재자/협조자”일 때 행위자〉개인〉직위직급명 행위자〉개인〉개인명[주2]
	/neo/DOC+대표기관코드(7)+년도(4)+일련번호(10).neo 〉 FileMetadata	문서일시 * 문서유형=“본문”일 때 문서제목, 문서행위자

이 가능함을 알 수 있고 주로 행위자 정보에 해당됨을 파악할 수 있다.

4.4 메타데이터 추출 알고리즘과 규칙

본 연구에서는 크게 3가지 추출 방식을 사용하였는데 첫번째는 데이터 필드 제목을 이용하는 방식이고 두번째는 데이터 필드 위치를 이용하는 방식이다. 마지막으로 세번째는 문단끝 부호를 이용하여 메타데이터를 추출하는 방식으로 추출 알고리즘을 구성하였다.

4.4.1 데이터 필드 제목 이용 방식

결재 문서의 일반기안문 서식은 고정적이며 필드의 위치가 변하는 않는 특징이 있다. 예를 들면 수신자에 이어 제목이 나타나며 제목 다음에는 본문 텍스트가 이어진다. 또 서식에는 변수에 해당되는 데이터 필드 자리만 고정되어 있는 필드가 있고 수신자와 문서제목처럼 필드 제목이 먼저 나타나는 필드가 있다.

필드 제목은 상수라고 하겠으며 바로 다음 나타날 필드 값을 예고하는 지표로 삼을 수 있다. 또한 다른 필드 제목과의 상대적인 위치를 이용함으로써 제목 없는 다른 필드를 찾는 지표로 삼을 수 있다.

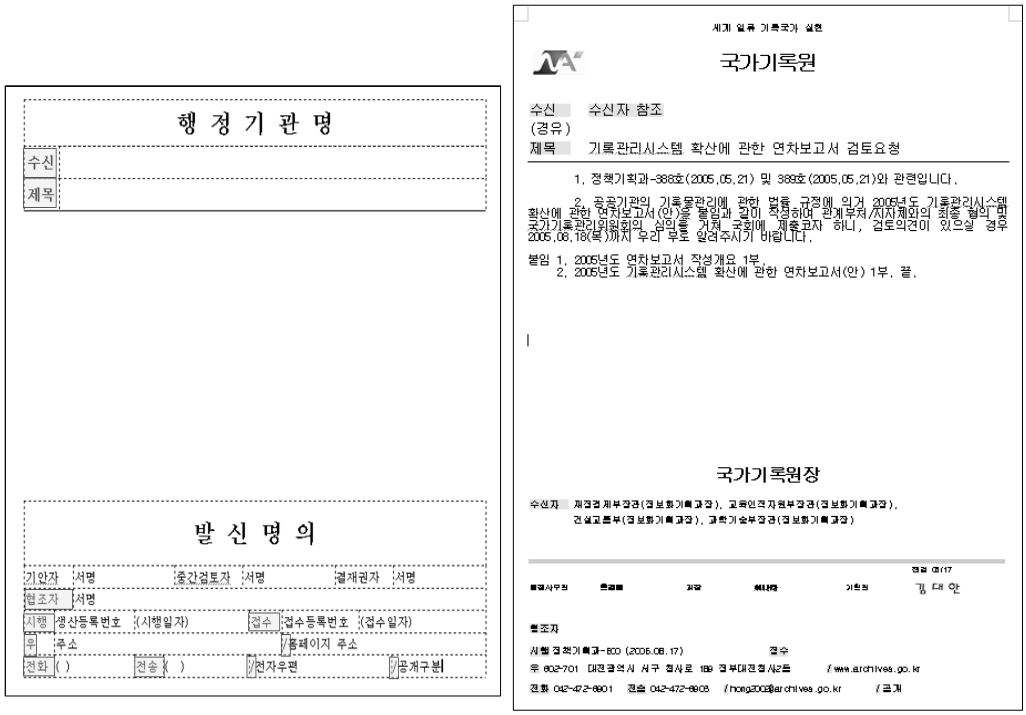
〈그림 10〉은 결재 문서에 필드 제목이 나타내어진 상태를 표시하였는데 필드 제목은 노란색으로 표시되어 있다. 이처럼 필드 제목이 있는 경우에는 바로 다음에 나타나는 값이 필드값이라고 판단하여 해당 필드에 해당하는 메타데이터를 추출하도록 하였다.

4.4.2 데이터 필드 위치 이용 방식

결재 문서의 일반기안문 서식은 고정적이며 필드의 위치가 변하지 않는 특징이 있는데, 일반기안문 양식 결재 문서 내 필드 순서를 살펴보면 〈그림 11〉과 같다.

수신자(②)에 이어 제목(③)이 나타나며 제목 다음 본문 텍스트가 뒤따른다. 본문 텍스트가 끝난 뒤 발신명의(④)가 나타나고, 수신자가 2인 이상일 경우 발신명의 다음에 수신자(⑤)가 나타나며 이후 결문의 데이터 필드들(⑥~㉓)이 순서대로 온다.

결문의 필드 순서 결재선은 왼쪽에서부터 기안자(⑨, ⑩), 검토자(⑪, ⑫)의 순서로 중간검토자가 2인 이상도 가능하며 협조자의 제목 필드인 "협조자"가 나올 때까지는 결재선이며 마지막이 최종 결재자(⑬, ⑭)이다. 결재선은 직위-서명의 쌍으로 분해되며 서명이 전자서명일 때는 이름 추출이 가능하다.



〈그림 10〉 결재 문서 필드 제목(Field titles in a document for approval)

〈그림 11〉 일반기안문 양식 결재 문서 내 필드 순서

마지막으로 결재일(⑧)은 최종 결재자의 서명 위에만 "월/일"의 형식으로 표시된다. 날짜 앞에 "전결", "대결"이 표시될 수 있는데, 결재일이 표시되는 줄은 직위-서명 쌍과 분리되며 서명란보다 위에 위치해 있기 때문에 직위-서명보다 먼저 추출되어야 하며 추출 시점에서는 결재선의 마지막 결재자 서명 위에 위치해 있는지 확인할 수 없다. 그러나 발신인명이나 수신자 다음이면서 시행일이나 접수일 전에 위치한 날짜 형식은 결재일이 유일하므로 이 구역에서 추출되는 MM/DD 형식의 값은 결재일로 간주할 수 있다.

4.4.3 문단끝 부호 이용 방식

결재 문서의 서식은 표로서 각 셀에 입력된 텍스트는 문단끝 부호((\0x0D0A)로 끝난다. 편집기에서 사용자가 직접 표에 문장을 작성한다면, 사용자가 임의로 한 셀 내에 여러 단락을 입력한 결과 여러 개의 문단끝 부호가 삽입될 수 있으나 전자문서시스템에 의해 표의 특정 셀에 데이터가 자동 삽입되는 경우에는 한 셀에 한 단락만 삽입된다. 따라서 문단끝 부호로 셀, 즉 데이터 필드를 구분할 수 있다.

4.4.4 데이터 추출 규칙

결재 문서 원문으로부터 서식을 무시하고 텍스트만 추출하면 〈표 9〉와 같이 문단끝 부호[7]를 포함하여 첫 문자부터 이어진 긴 문자열을 얻게 된다. 이 문자열로부터 〈표 10〉의 규칙을

7) 문단끝부호: $ (0x0D0A).

〈표 9〉 일반기안문 양식 결재 문서의 추출 텍스트 샘플

세계 일류 기록국가 실현 국가기록원 수신 수신자 참조$(경유)$제목$기록관리시스템 확산에 관한 연차보고서 검토요청$1. 정책기획과-388호(2005.05.21) 및 389호(2005.05.21)와 관련입니다. $2. 공공기관의 기록물관리에 관한 법률 규정에 의거 2005년도 기록관리시스템 확산에 관한 연차보고서(안)를 붙임과 같이 작성하여 관계부처/지자체와의 최종 협의 및 국가기록관리위원회의 심의를 거쳐 국회에 제출코자 하니, 검토의견이 있으실 경우 2005.08.18(목)까지 우리 부로 알려주시기 바랍니다.$붙임 1. 2005년도 연차보고서 작성개요 1부.$2. 2005년도 기록관리시스템 확산에 관한 연차보고서(안) 1부. 끝.$국가기록원장$수신자 재정경제부장관(정보화기획과장), 교육인적자원부장관(정보화기획과장),$건설교통부(정보화기획과장), 과학기술부장관(정보화기획과장)$전결 08/17$행정사무관 홍길동 과장 이나라 기획관$협조자 시행 정책기획과-500 (2005.08.17) 접수$우 302-701 대전광역시 서구 청사로 189 정부대전청사 2동 /www.archives.go.kr$전화 042-472-3901 전송 042-472-3906 /hong2002@archives.go.kr /공개

노출된 개인정보, 멍드는 신용사회 $ ○○○○부 $ $ $수신 $○○○○부(복지담당관) $(경유) $ $제목 $설문조사를 위한 소속 공무원 연락처 보고 $ $ $ 1. ○○○○부 복지담당관-394(2004.11.16)호와 관련입니다. $ 2. 우리부 소속 공무원 연락처 붙임과 같이 보고합니다. $ $붙임 설문조사용 소속 공무원 연락처 1부. 끝. $ $ $ $ $ $ $ $ $ $ $ $ $ $ $ $○○○○부장관 $ $ $수신자 $ $ $ $ $ $ $ $전결 10/20 $ $ $담당자 $김미주 $총무팀장 $이순철 $관리과장 $강산애 $ $ $협조자 $ $ $ $ $ $ $ $ $시행 $관리과-4432 (2004.09.20.) $접수 $복지담당관-189 (2004.11.20.) $우 $135-222 $○○시 ○○로13(○○동 91-2) $/ $ $전화 $070-111-2222 $전송 $ $/ $kim70@hotmail.com $/ $공개 $ $ $ $

〈표 10〉 일반기안문 양식 결재 문서의 데이터 필드 추출 규칙

데이터 필드	추출 규칙
수신자	"수신" 다음에 나오는 "제목" 사이 문자열 또는 "수신" 다음 문단끝 부호까지의 문자열 * 구한 문자열이 "수신자 참조"이면 결문에서 다시 수신자 추출 * 구한 문자열이 "내부결재"이면 수신자/발신자 불필요
제목	"제목"과 문단끝 부호 사이 문자열
발신명의	생산기관명으로 시작하고 단어 끝이 "장관", "장", "사장"으로 끝나는 문자열
수신자	두 번째 나오는 "수신자" 이후 문단끝 부호 사이 문자열+"-" 다음의 숫자
기안자	발신명의 다음 또는 수신자 다음 "전결"이나 "대결" 또는 mm/dd 형식 숫자가 아닌 첫 번째 단어부터 "협조자" 전까지 문자열을 단어 단위로 추출한 뒤 2단어씩 쌍으로 할 때 첫 번째 쌍(서명은 반드시 전자서명은 아니므로 이미지 서명일 경우를 고려하여 추출해야 함) 앞 단어는 기안자 직위, 뒤 단어는 기안자의 서명. 전자서명일 경우에만 이름은 추출 가능
중간검토자	위에서 얻은 쌍 중 첫 번째와 마지막 쌍을 제외한 중간 쌍은 검토자 직위와 서명에 해당. 자서명일 경우에만 이름은 추출 가능
결재자	위에서 얻은 쌍 중 마지막 쌍은 검토자 직위와 서명에 해당. 자서명일 경우에만 이름은 추출 가능
협조자	"협조자"에 이어 "시행" 사이 문자열. 공백이거나 또는 단어 2개로 이루어진 쌍이 반복됨. 첫 번째 단어는 직위, 두 번째 단어는 전자서명일 경우 이름
문서번호	"시행" 또는 "접수"와 "(" 사이 한글 또는 숫자 문자열
시행/접수일자	"시행" 또는 "접수"를 따라오는 문서번호 패턴에 이어 오는 괄호 안의 yyyy.mm.dd 형식의 문자열
주소	"우"에 이은 우편번호 패턴([0-9]{3}-[0-9]{3}) 다음부터 "/" 사이 문자열
전자우편	"/"를 이어 나오는 이메일 패턴
공개구분	"/"를 이어 나오는 "공개" 또는 비공개(n, n..) 또는 부분공개(n, n...) 패턴의 문자열

복합적으로 적용하여 필드별 값인 메타데이터를 추출할 수 있다.

4.5 전자기록물 메타데이터 추출 및 검증 기술 구현 결과[8)]

4.5.1 구현 내용

본 연구에서 개발한 시스템은 첨부 파일 본문에서 5개 항목(기안자, 기안일자, 제목, 최종결제자, 시행일)을 추출하여 비교 검증하도록 구현하였다.

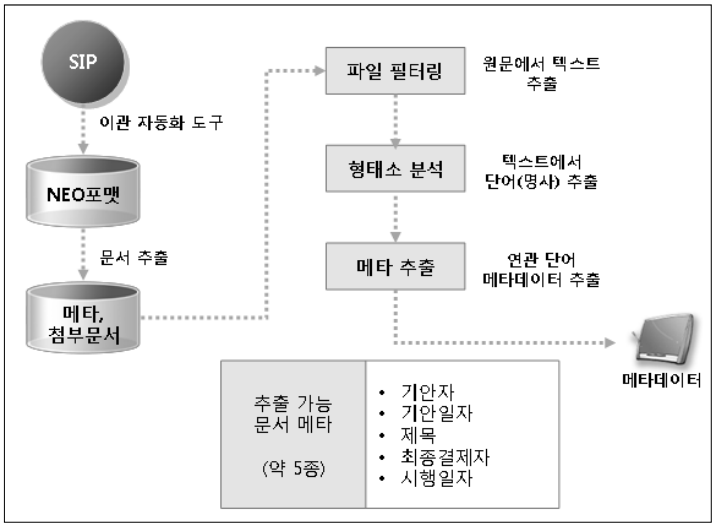
본문의 다양한 필드에서 추출할 수 있는 데이터 중에서 영구보존기록물의 메타데이터로 활용할 수 있는 필드값만을 추출하였다. 또한 본문을 제외한 첨부 파일은 정형화된 양식을 갖는 파일 형태가 아니므로 메타데이터를 추출하지 않았다.

원문으로부터 메타데이터를 추출과는 과정은 〈그림 12〉에 표시하였는데, 원문 파일로부터 텍스트를 추출한 후 형태소 분석을 수행하여 명사를 추출한 다음 메타데이터 항목에 해당되는 필드를 찾아 데이터를 추출한다.

테스트 시스템에서는 추출된 메타데이터와 이관된 메타데이터와의 비교 검증을 수행하여 상호 일치하지 않는 경우 메타데이터를 수정할 수 있도록 구현하였다.

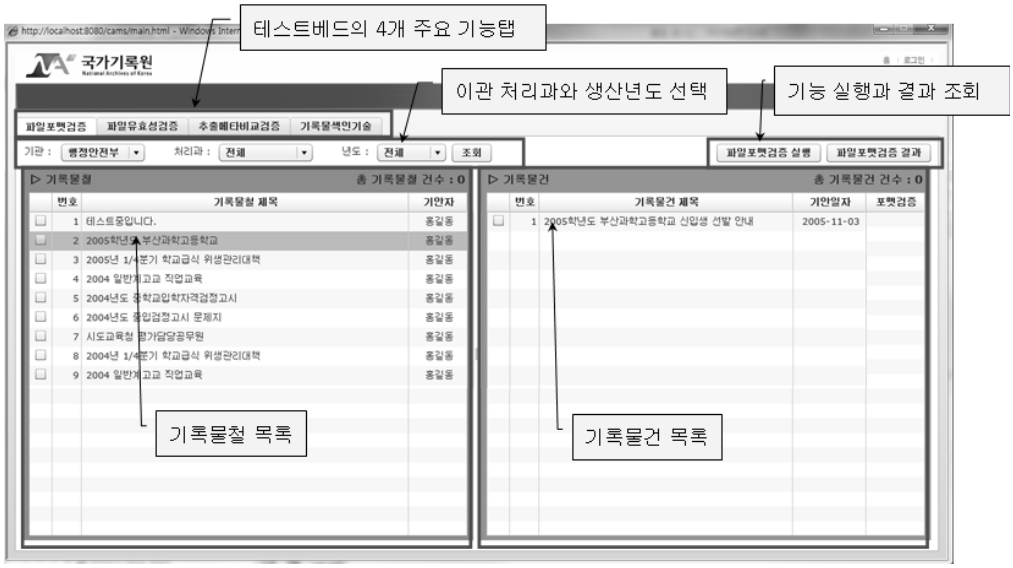
4.5.2 구현 화면 결과

전자기록물 메타데이터 추출 및 검증 기능의 사용자 웹 인터페이스는 〈그림 13〉의 테스트베드 메인화면에서와 같이 주요 4개 기능별 탭 중에서 3번째 탭인 "추출메타비교검증"으로 구성하였다.



〈그림 12〉 메타데이터 추출 과정

8) "전자기록물 검증 기술 및 차세대 그린 전자기록관리 체계 인프라 응용 기술 연구" 완료보고서(국가기록원 2010a)를 토대로 작성됨.



〈그림 13〉 테스트베드 메인 화면(Main page for the test-bed)

〈그림 14〉에서와 같이 "추출메타비교검증" 탭에서 기록물철과 기록물건을 선택한 뒤 '추출메타비교검증 실행' 버튼을 클릭하여 메타데이터 추출 검증을 실행하면 각 기록물건의 전자파일로부터 메타데이터를 추출한다.

메타데이터 추출을 실행한 후 〈그림 15〉와

〈그림 14〉 메타데이터 추출 검증 화면

〈그림 15〉 메타데이터 추출 검증 결과 화면

같이 '추출메타비교검증 결과' 버튼을 클릭하면 검증 결과를 조회할 수 있다. 해당 기록물건에 대하여 원문 파일로부터 추출한 메타데이터와 이관데이터로 전송된 메타데이터를 나란히 조회할 수 있으며, 비교 후 오류를 수정하려면 '오류 수정' 버튼을 클릭하여 추출한 메타데이터로 변경할 수 있다.

그러나, 이관된 메타데이터를 수정하는 작업은 기술적인 면과 정책적인 면을 함께 고려하여야 하며 본 연구에서는 기술적 가능성만을 테스트하였다.

5. 결론 및 제언

본 연구에서는 인수 전자기록물에 대하여 원문으로부터 메타데이터를 추출하고 이를 이관된 메타데이터와 비교 검증하는 기술적인 면을 검토하고 테스트 시스템을 개발하여 그 적용 가능성을 모색함으로써 2015년 대량 전자기록물 인수에 대비하고자 하였다.

즉, 영구기록관리시스템으로 이관되는 기록에는 기록물철과 기록물건이 있으나 본 연구에서는 기록물건의 첨부파일과 같은 경우는 특정한 양식을 가지고 있지 않기 때문에, 본체에 해당되는 본문전자파일의 메타데이터를 검증하기 위한 각종 요소 기술을 분석하고 검증 기능 및 체계를 수립하였다.

본 연구를 통해 개발된 '메타데이터 추출 및 검증 도구'는 2015년 대량·대용량 기록물 이관을 대비한 CAMS 시스템 개발 사업에서 기록물 검증 단계에 적용 가능하도록 테스트 시스템으로 개발하였으며 차기 사업에 반영하여 성능 및 기능 향상이 가능하도록 구성하였다.

따라서 본 연구를 통해 개발된 "메타데이터 추출 및 검증 도구"를 대량·대용량 기록물 이관 사업의 검증 단계에 적용하여 이관되는 메타데이터에 대한 검증에 활용할 수 있다.

마지막으로 본 연구를 통해 원문을 활용한 메타데이터 검증에 필요한 핵심 기술을 축적하여 체계를 수립하였는바, 이러한 성과는 향후 영구기록관리시스템 개선 사업에 적용하여 2015년 대량 기록물 본문파일에서 추출한 메타데이터를 이용하여 인수 시 검수 자동화를 구현할 수 있으며, 기타 기록관리시스템은 물론 기록생산시스템에서도 메타데이터 검증 자동화 프로세스를 적용하여 기록물 생산과 보존의 오류를 방지하는데 활용할 수 있다.

참 고 문 헌

강승식. 2004. 한글 문서의 색인어와 색인 기법. 『정보과학회지』, 22(4): 72-77.

국가기록원. 2010a. 전자기록물 검증 기술 및 차세대 그린 전자기록관리 체계 인프라 응용 기술 연구 완료보고서.

_____. 2010b. 차세대 전자기록관리 인프라 연구 개발 연구보고서.

권순만 외. 2004. 단어기반 웹문서 검색을 위한 효과적인 단어 가중치의 계산. 『한국정보과학회 2004년도 가을 학술발표논문집』, 31(2): 169-171.

김남희. 2005. 국회도서관 시소러스 구축과 활용 그리고 유지관리. 『국회도서관보』, 42(11)(통권 제319호): 36-49.

김태중. 2003. 시소러스에 관한 일반적 고찰. 『국회도서관보』, 40(3)(통권 제289호): 40-49.

류계자. 2001. 정보검색을 위한 형태소 분석기의 기능 확장에 관한 연구. 한양대학교 산업대학원. 32-35.

이재윤. 2003. 역문헌빈도 가중치의 재검토. 한국정보관리학회. 『2003년도 제10회 학술대회 논문집』, 253-261.

최호철. 2003. 특수분야 및 띄어쓰기 오류 문서 분석을 개선한 형태소 분석기의 구현. 중앙대학교 정보정보대학. 19-20.

한상길. 1994. 시소러스를 이용한 신문기사 데이터베이스 색인시스템에 관한 연구. 『정보관리학회지』, 11(1): 125-144.

Evaluation of characterisation tools Part 1: Identification, Johan van der Knijff, Carl Wilson, p.4.

〈https://bytebucket.org/jhove2/main/wiki/documents/JHOVE2-functional-requirements-v1_4.pdf〉, p.6.

J. H. Lee and J. S. Ahn. 1996. Using N-Grams for Korean Text Retrieval. ACM SIGIR Conference on Research and Development in Information Retrieval, 216-224.

Larry Stone. 2008. BitstreamFormat Renovation: DSpace Gets Real Technical Metadata. Open Repositories Conference 2008.

Medelyan, O. 2005. "Automatic Keyphrase Indexing with a Domain-Specific Thesaurus." Master Thesis. University of Freiburg, Germany.

Medelyan, O. and I. H. Witten. 2005 "Thesaurus-based index term extraction for agricultural documents." In: Proc. of the 6th Agricultural Ontology Service (AOS) workshop at EFITA/WCCA 2005, Vila Real, Portugal.

Microsoft Office File Format Documents. 〈http://msdn.microsoft.com/en-us/library/cc313105(office.12).aspx〉.

P. M. Roget. 1852. Thesaurus of English Word and Phrase.

SK C&C. 2010. 차세대 전자기록관리 인프라 연구 개발. p.50.